Experimental Results: Random digits and hypothesis testing

---

48 marks

In this question, we will explore the data collected in class.

You will need to download the data `class_data.csv` from the assignment website and save it somewhere. Supposing the data to have been saved in a directory `dataDirectory`, read it into R as

```
# Usual helper for paths
path_concat <- function(path1, path2, sep="/") {
                        paste(path1, path2, sep = sep)
}

data <- read.csv(file = path_concat(dataDirectory, "class_data.csv"))
```

Having loaded the data, you might have a look at its contents using any of the standard functions:

```
# just print it
data
# view its contents as a spreadsheet (in RStudio)
View(data)
# or, look at its data structure
str(data)
```

You will find that it is a `data.frame` with variables

```
names(data)
```

```
## [1] "random_digit"  "student_digit" "green_card1"   "green_card2"
## [5] "red_card1"     "red_card2"
```

Each row contains one student's answer to the questions associated with these names.

Here we will only consider the results on the digits (0-9) obtained in class.

Recall how the data on the digits 0, 1, …, 9 were collected.

Each person was first asked to write the words "random digit" on a card. They were then given a few seconds to think up a single *random* digit from 0 to 9 and then record it on the card.

On the other side of the card, each person then wrote "student digit" and below it recorded the *last* digit of their student id number.

These two digits provide the values for the variables `random_digit` and `student_digit` appearing in the data set `data` above.

---

1. Suppose a digit, d, is generated as a realization from a random variable D which is uniformly distributed on the digits $\{0, 1, 2, 3, \ldots, 8, 9\}$. That is, for any value $d \in \{0, 1, 2, 3, \ldots, 8, 9\}$ we have

$$Pr(D = d) = \frac{1}{10}.$$

   a. *(1 mark)* Determine the expectation $E(D)$.

   b. *(2 marks)* Determine the expectation $E(D^2)$ and hence the standard deviation $SD(D)$.

   c. *(1 mark)* Determine the median of D.

   d. Suppose we consider the random variable C which for some fixed value $d \in \{0, 1, 2, 3, \ldots, 8, 9\}$ takes values

$$C = \begin{cases} 1 & \text{when } D = d \\ 0 & \text{when } D \neq d \end{cases}$$

   which implies

$$Pr(C = 1) = Pr(D = d) = \frac{1}{10}$$

   and

$$Pr(C = 0) = Pr(D \neq d) = 1 - Pr(D = d) = \frac{9}{10}.$$

   Suppose we have $C_1, C_2, \ldots, C_n$ independent and identically distributed random variables with the same distribution as C.

   Let $X = \sum_{i=1}^{n} C_i$.

      i. *(1 mark)* What is the name of the probability distribution of X?

      ii. *(1 mark)* Hence, write down an expression $Pr(X = x)$ for $x \in 0, 1, \ldots, n$.

      iii. *(1 mark)* Hence, write down an expression $E(X)$.

   e. Using the `data` from class, for **each** of the variables `random_digit` and `student_digit`, calculate (show your code in each case) its

      i. *(2 marks)* sample average,

      ii. *(2 marks)* sample standard deviation,

      iii. *(2 marks)* sample median,

      iv. *(3 marks)* and compare these to the corresponding theoretical values from the distribution of D.

      v. *(1 mark)* Which of `random_digit` and `student_digit` have sample values closer to the theoretical values.

   f. *(3 marks)* Calculate $Pr(X = x)$ when $n = 42$ and $x = 0$, 5, and 10.

---

2. We are interested testing the hypothesis

   H: the observed digits d_1, d_2, ..., d_n are independent realizations of a random variable D uniformly distributed on the digits $\{0, 1, 2, \ldots, 9\}$.

   In particular, we are interested in testing this hypothesis for each of two samples `student_digit` and `random_digit`.

   a. *(4 marks)* The function `stem()` is a simple way to get a quick picture (a "stem and leaf plot") of the distribution of a set of digits. Use `stem()` to construct a picture of each of the following:

      i. the values of `student_digit`,
      ii. the values of `random_digit`,
      iii. and, for comparison, a sample of the same size from a uniform distribution on the digits using the function `sample()`.

      Which of `student_digit` or `random_digit` looks more like it might have come from a Uniform on the digits? Why?

   b. A more formal way to assess whether a sample of values appear to come from a hypothesized distribution is to calculate the *Pearson's chi-squared test* of "goodness of fit" statistic. This is generally expressed as

$$\mathrm{X}^2 = \sum_{i=1}^{m} \frac{(o_i - e_i)^2}{e_i}$$

      where o_i is the observed number of values in the ith "cell", e_i is the expected number to fall into that cell according to the hypothesized model, and m is the total number of (non-overlapping) cells.

      In our case, the cells are the 10 different possible digits (so $m = 10$) and o_i is the number of digits d_1, d_2, ..., d_n equal to i, for each $i \in \{0, 1, \ldots, 9\}$. The values e_i are the *expected* number of digits equal to i for each $i \in \{0, 1, \ldots, 9\}$, when the hypothesized model is true. In this case $e_i = n/m$ for all i.

      X^2 is a discrepancy measure which is larger whenever the o_i are relatively far from their expectation under the model, e_i. The larger is X^2, the greater is the evidence against the model.

      If the hypothesized model is true, then the distribution of X^2 can usually be approximated as a χ^2_k distribution having degrees of freedom $k = m - \#\text{constraints on the model}$. In our case, there is only one constraint (the total of the expected values must sum to n; $\sum_{i=1}^m e_i = n$). So here, the degrees of freedom are $k = m - 1$. The usual rule-of-thumb is that χ^2_k is a good approximation of the distribution of X^2 provided $e_i \geq 5$ for all $i = 1, \ldots, m$.

      i. *(4 marks)* Write a function `count_digits()` which takes a vector of digits `d` and returns a numeric vector whose ith element contains the number of values in `d` which were equal to $i - 1$. That is, write

```r
count_digits <- function (d) {
       # Your code here
       }
# for example if
my_digits <- c(0, 1, 3, 4, 7, 1, 4, 9, 7, 4)
# then
count_digits(my_digits)
# would return a vector equal to
c(1, 2, 0, 1, 3, 0, 0, 2, 0, 1)
```

         Note, that we will assume that `d` will be an integer vector containing only values in $\{0, 1, 2, \ldots, 9\}$.

No error checking is required for now.

ii. *(2 marks)* Demonstrate your function on the digits of the variable `student_digit` and on the digits of the variable `student_digit`.

iii. *(4 marks)* Write the function `Pearson_chi_sq(observed, expected)` which calculates X^2 where `observed` is a vector of observed counts, and `expected` is a vector of expected counts given the model. The vector `expected` should have the same length as `observed` or be of length 1.

Again, for expediency it will be assumed that both `observed` and `expected` vectors contain only integer elements in $\{0, 1, 2, \ldots, 9\}$.

However, you should check that the lengths of `observed` and `expected` match and stop if they do not. If the length of `expected` is only 1, then create a vector of length equal to that of `observed` with the value of `expected` repeated.

```
Pearson_chi_sq <- function (observed,
                            expected = sum(observed)/length(observed)) {
   # Your code here;  No, you cannot use the function chisq.test()
   }
```

iv. *(2 marks)* Check your function by comparing the values of `Pearson_chi_sq(observed)` to results of `chisq.test(observed)$statistic` for `observed` being each of

`count_digits(data$student_digit)` and `count_digits(data$random_digit)` in turn.

v. *(2 marks)* Using the function `pchisq()` calculate the p-value testing the uniformity hypothesis using your calculated `Pearson_chi_sq()` value for the counts of the digits in each of the `data` variables `student_digit` and `random_digit`.

Show that your p-values agree with those produced by `chisq.test(observed)$p.value` for the counts of the digits in each of the `data` variables `student_digit` and `random_digit` as `observed`.

vi. *(3 marks)* Rather than depend upon the validity of the χ^2_k approximation, we could *simulate* the distribution of X^2 by calculating its value on B samples of size `n = nrow(data)` = 42 generated by the function `sample()`.

Using the function `sapply()` and the function `sample()`, together with your functions `Pearson_chi_sq()` and `count_digits()` to create a function `get_chisqs(n, B = 1000)` that generates the chi-squared statistic on each of `B` independently drawn samples, each being of size `n` from a uniform distributon on the digits $\{0, 1, \ldots, 9\}$, and returns a numeric vector of length `B` containing the `B` chi-squared statistics.

That is, write

```
get_chisqs <- function (n, B = 1000) {
   # Your code here
}
n <- nrow(data)
results <- get_chisqs(n = n, B = 1000)
```

vii. *(3 marks)* Use your function `get_chisqs()` to get `B = 10000` independent pseudo-random realizations of the X^2 statistic from a sample of size `n = nrow(data)` = 42 from the uniform distribution on the digits.

That is, execute the following (N.B. in RMarkdown change header to eval = TRUE):

```
n <- nrow(data)
B <- 10000 # TEN thousand
set.seed(314159) # So we all get the same values
```

```r
chisq_stats <- get_chisqs(n = n, B = B)
hist(chisq_stats,
     col = "lightgrey",
     main = "Simulated Pearson test null distribution",
     xlab = "test stat")
```

To this histogram, add a vertical line in "red" where the corresponding statistics you calculated should be for the `student_digit` variable, and add a vertical line in "blue" where the corresponding statistics you calculated should be for the `random_digit` variable.

Put a legend in the top right that identifies the lines.

Based on this histogram, which collection of digits seem less likely to have been generated as a random sample from a uniform distribution of digits? Why?

viii. *(2 marks)* The simulated distributions can be used to calculate approximate p-values by simply evaluating the proportion of the simulated test statistics which are greater than or equal to the X^2 values for each of the counts corresponding to `student_digit` and `random_digit`.

Calculate these two p-values using the simulated test null distribution given by the vector `chisq_stats` above. Show your code.

ix. *(2 marks)* What do you conclude about the two hypotheses?